Audio Engineering Society

Convention Express Paper 295

Presented at the AES 157th Convention
2024 October 8-10, New York, NY, USA

This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Headphones vs. Loudspeakers: Finding the More Effective Monitoring System

William Shehan, Wesley Bulla, Song Hui Chon, and Doyuen Ko

M.S. Audio Engineering, Mike Curb College of Entertainment and Music Business, Belmont University, Nashville, TN 37212

Correspondence should be addressed to William Shehan (will.shehan@bruins.belmont.edu)

ABSTRACT

There has been debate over whether headphones or loudspeakers are the more effective monitoring system. However, little discussion has been conducted regarding the effectiveness of the two systems in the context of mixing and music production tasks. This study aimed to examine how the two monitoring systems would influence users' performance in production tasks. An experiment was designed in which the subject was asked to match the boost level of a given frequency band in a reference recording until the boost sounded identical to the reference. A group of six audio engineering graduate students completed the experiment using headphones and loudspeakers. Subjects' adjustment levels across eight frequency bands were collected. Results suggest that both monitoring systems were equally effective for energy matching in specific frequency bands. Future research is called for with more subjects with diverse levels of expertise in various production-oriented tasks to better understand the potential effect of each playback system.

1 Introduction

There is an ongoing question of which monitoring system, headphones or loudspeakers, is most beneficial to users in music production. Prior reports appear to support the superior effectiveness of loudspeakers [1]. However, past studies focused on listening tasks such as user preferences [2][3] and detection thresholds [4] and therefore inferred system performance via passive reporting. To better understand the influence of each playback system, performance should be tested with tasks directly associated with production-oriented skills like applying equalization (EQ), adding reverberation, and blending ambiance effects. This study aims to determine whether loudspeaker monitoring is the best option or if headphones could offer a more effective alternative.

2 Literature Review

Previous studies have compared headphones and loudspeakers when completing listening tasks. King et al. [2] examined the preferred listening levels of mixing engineers when using headphones and loudspeakers. In two genres tested (classical and rock), the average preferred listening level for mixing was lower with loudspeakers than with headphones. This result suggests that a lower volume on the loudspeakers provided adequate detail when mixing in those genres, but also that the headphones needed to be turned up to reveal that same detail. This could support the idea that loudspeakers would be more helpful to the engineer during listening tasks. Similarly, a test by Olive et al. [3], conducted with both trained and untrained listeners, found their preferred parameters when using headphones and loudspeakers. The frequency response was flattened for both devices beforehand. The results suggested

that the listeners' preferred response for loudspeakers was similar to that for headphones. This could in turn imply that there is minimal difference in listeners' frequency response preferences between the two monitoring systems. Although these studies reveal information about user preference when using headphones and loudspeakers, they do not compare the listeners' performance with the two systems.

In another test by Choisel and Martin [4], detection thresholds were found for phase mismatch in noise from 63 Hz to 8000 Hz. Results of this study indicated that these thresholds were reasonably consistent when using loudspeakers, whereas they were more frequency-dependent when using headphones. This suggests that loudspeakers could provide a more even detection threshold for minute changes, supporting the idea that loudspeakers could be a more consistent choice for producers in a working environment. It is necessary, however, to conduct a new experiment with easily measurable results in which subjects are tested on production tasks with the two systems.

3 Methods

An experiment was conducted in a semi-anechoic chamber, where subjects were asked to match the target boost level of a given frequency band within an audio clip. The anechoic room was selected because it minimizes the influence of the acoustic environment on the listener's decision.

To ensure that stimuli reflected the professional quality of modern recordings, eight 30-second audio clips [5] were chosen from Billboard's Top 20 Songs listed on a randomly chosen date [6]. Table 1 below lists the eight songs used in this study, along with their assigned clip number and respective genre.

For each song, eight "stem" tracks were created for eight frequency bands of interest (100 Hz, 200 Hz, 400 Hz, 800 Hz, 1000 Hz, 2000 Hz, 3000 Hz, and 4000 Hz). The range of frequencies was chosen to represent the range within which most of the frequency content of the songs would be present. The frequency content above 4000 Hz and below 100 Hz was less consistent among all the tested songs, which could have led to disproportionate levels of response in those frequencies from the subjects.

Clip #	Song Title	Artist	Genre
01	"About Damn Time"	Lizzo	Pop
02	"As It Was"	Harry Styles	Pop
03	"Wasted On You"	Morgan Wallen	Country
04	"First Class"	Jack Harlow	Hip-Hop/Rap
05	"Wait For U"	Future (feat. Drake)	Hip-Hop/Rap
06	"BREAK MY SOUL"	Beyonce	Pop
07	"Heat Waves"	Glass Animals	Pop
08	"Late Night Talking"	Harry Styles	Pop

Table 1. A list of the eight songs from which the specified 30-second audio clips were taken. All the songs above were chosen from the Billboard Top 20 on July 30, 2022.

Each stem track was made by processing the original stereophonic recording through a bandpass filter at a 48 dB/octave slope, isolating the frequency of interest. Using the same filter, a "bed" track was also created for each song. It contained the frequency contents of the original audio clip except for the frequency range of interest. Together, the bed and stem tracks represent the original sound of the clip. A boost amount between 1.0 dB and 10 dB, in 1.0 dB increments, was added to a selected stem to create a reference track for each song. These fader settings were then fixed and routed to the "A" output. Subjects had a duplicate set of stems to manipulate. These faders were routed to the "B" output. The audio clip, the frequency band of interest, and the boost amount were counterbalanced and randomly selected for each trial. There was only one band of interest (given) for subjects to focus on for each trial. Subjects could switch between their output and the reference output via an A-B switch and alter their level of EQ boost until the two recordings sounded identical. There was no numerical feedback given to subjects as they adjusted the fader. Six graduate students studying audio engineering completed two sets each of eight trials using headphones (HP) and loudspeakers (LS) for each target EQ. Subjects' performance was measured as the absolute value (in dB) of the difference between the target boost level and the user's chosen EQ level [7].

4 Results

A non-parametric analysis of variance (ANOVA) revealed that one subject's performance (M = 3.97, SD = 6.43, N = 31) did not agree with the rest of the population (M = 2.33, SD = 2.33, N = 159), $X^2(N = 31) = 11.1$, $p = .05$, as illustrated in Fig. 1. The data from this particular subject were therefore omitted from the analysis.

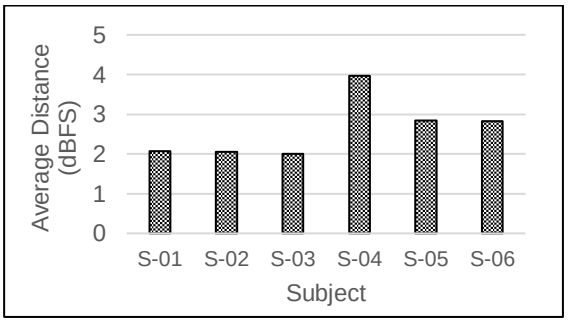


Fig. 1: The average distance per subject between the correct boost level and the subject-chosen level.

As shown in Table 2, individual subjects' mean differences ranged from 1.84 dB to 3.00 dB (SD = 0.52 dB) for the HP condition and 1.34 dB to 2.85 dB (SD = 0.58 dB) for the LS condition. This presented a similar standard deviation across subjects for both systems, implying that the headphone test results did not have much more variation than the loudspeaker results. Comparative subject performance between HP and LS conditions was calculated by taking an absolute difference between each subject's means during the HP condition and the LS condition. These values are listed in the far-right column of Table 2, ranging from 0.05 dB to 1.48 dB (SD = 0.58 dB). The standard deviation of 0.58 dB for this column shows a similar level of variability to those of the individual conditions.

Subject	Headphones		Loudspeakers		HP vs. LS
	M	*SD*	*M*	*SD*	
S-01	2.81	3.61	1.34	0.97	1.48
S-02	2.06	2.19	1.94	1.53	0.12
S-03	1.84	1.43	2.18	2.72	0.34
S-05	3	3.31	2.52	1.47	0.48
S-06	2.8	2.32	2.85	2.35	0.05
M	2.5	2.57	2.16	1.81	0.49
SD	0.52		0.58		0.58

Table 2. Mean (M) and standard deviation (SD) of each subject's adjustment differences using headphones and loudspeakers.

As presented in Fig. 2, Fig. 3, and Fig. 4, respectively, no significant influence of the audio clip ($\chi^2(N = 19) = 10.11$, $p = .18$), frequency band ($\chi^2(N = 19) = 8.11$, $p = .32$), or boost amount ($\chi^2(N = 8) = 7.85$, $p = .64$) on overall performance scores were found. This means that these variables did not significantly impact the subjects' responses, which in turn suggests that, if any significance was found from further analyses, it would be from the effect of the monitoring system.

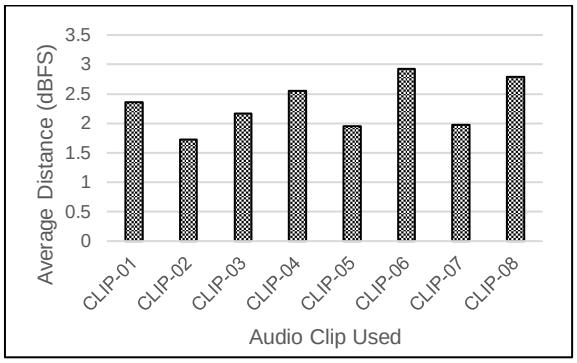


Fig. 2: The average distance from the correct boost level per audio clip across all subjects.

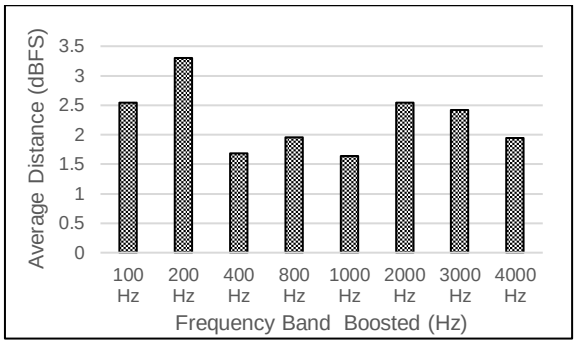


Fig. 3: The average distance from the correct boost level per frequency band across all subjects.

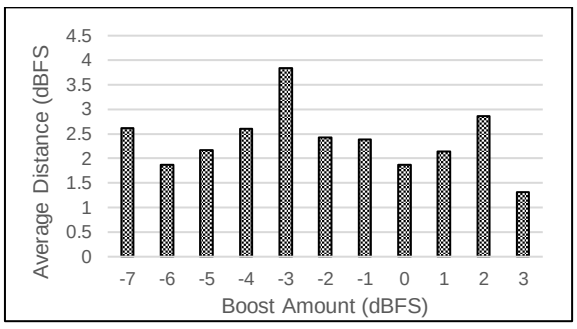


Fig. 4: The average distance from the correct boost level for each boost amount across all subjects.

A Wilcoxon matched pairs test could not reject the null hypothesis for a difference between monitoring systems, LS: $M = 2.16$, $SD = 0.58$; HP: $M = 2.50$, $SD = 0.52$ ($Z = –0.25$, $N = 79$, $p = .81$). This means that, overall, the listeners performed about the same when using the two systems. Results suggest that neither monitoring system was more advantageous when subjects were adjusting EQ levels to match the level of a given reference.

5 Discussion

Initial ANOVAs (by Subject, Clip, Band, and Boost) examined a possible significant effect of these parameters on the results. Through this step, one subject was found to be strongly skewing the data. This particular subject's data were different from the rest of the subjects being tested, and therefore were excluded from further analyses. In future testing, this same analysis will need to be conducted to identify inconsistent skill levels within the subject pool. All other variables in ANOVAs turned out to be not significant, which implies that they did not impact the overall results. This means, for example, that none of the eight audio clips affected the results more than any others. The relative similarity in results among all clips can be seen in Fig. 2, which shows the average distance resulting from each clip used in the study.

The final statistical test, the Wilcoxon Matched Pairs Test, found the effect of monitoring system to be not significant in this experimental scenario. What can be extrapolated from this is that the use of headphones or loudspeakers on this task would lead to largely similar results. With this information in mind, producers could simply choose the monitoring system which they are most comfortable working with, or whichever is most convenient for the situation. However, since these results are dependent on our experimental situation, more production-oriented tests will need to be conducted in order to verify this conclusion.

6 Conclusions

The purpose of this study was to explore the influence of modality (headphones and loudspeakers) on subject performance in production-oriented tasks. Both systems showed comparable subject performances. The results suggest that, in a production-oriented context similar to the experimental scenario, whichever system the engineer or producer prefers will be equally acceptable for the task. However, these results should be understood with caution, considering the variety of skill levels and the small number of subjects in this study. Further studies should include subjects with diverse training and incorporate additional production-oriented measures like the use of reverberation and ambiance effects. In doing so, the effect of the playback system may be parsed and quantified in light of demonstrated skill level, allowing future research to provide a more detailed understanding of the effect of the two playback systems. In addition, efforts should be made to examine frequencies that were not observed in the present experiment, especially those above 4000 Hz, to examine the impact of modality in high frequencies.

References

[1] F. Toole, "The measurement and Calibration of Sound Reproducing Systems," *J. Audio Eng. Soc*., vol. 63, no. 7/8, pp. 512-541, (DOI: http://dx.doi.org/10.17743/jaes.2015.0064), 2015.

[2] R. King, B. Leonard, and G. Sikora, "The Effects of Monitoring Systems on Balance Preference: A comparative study of mixing on headphones versus loudspeakers," in *Audio Engineering Society 131st Convention*, New York, NY [https://www-aes-org.bunchproxy.idm.oclc.org/e-lib/browse.cfm?elib=16091], Oct. 2011.

[3] S. Olive, T. Welti, and E. McMullen, "Listener Preferences for In-Room Loudspeaker and Headphone Target Responses," in *Audio Engineering Society 135th Convention*, New York, NY [https://www-aes-org.bunchproxy.idm.oclc.org/e-lib/online/browse.cfm?elib=17042], Oct. 2013.

[4] S. Choisel and G. Martin, "Audibility of phase response differences in a stereo playback system. Part 2: Narrowband stimuli in headphones and loudspeakers," in *Audio Engineering Society 125th Convention*, San Francisco, CA [https://www-aes-org.bunchproxy.idm.oclc.org/e-lib/browse.cfm?elib=14711], Oct. 2008.

[5] Apple Inc, "Apple Music: Lizzo, "About Damn Time," Harry Styles, "As It Was," Jack Harlow, "First Class," Future, "Wait For U," Beyonce, "Break My Soul," Glass Animals, "Heat Waves," Harry Styles, "Late Night

Talking," Morgan Wallen, "Wasted on You"," [Online]. Available: https://music.apple.com/us/browse.

[6] Billboard, "Billboard Hot 100," Penske Media Corporation, Billboard Media, LLC, billboard.com, 30 July 2022. [Online]. Available: https://www.billboard.com/charts/hot-100/2022-07-30/. [Accessed 14 February 2024].

[7] S. Bech and N. Zacharov, Perceptual audio evaluation - Theory, method and application., Chichester, England: Wiley (DOI: http://dx.doi.org/10.1002/9780470869253), pp. 29-38, 2006.